Data and Services Discovery projects - Transformative Data Collections

Title

Supporting FAIR machine maintenance data for Industry and Researchers in the Centre for Transforming Maintenance through Data Science

Approach

Identify security requirements and barriers to working with industry and sourcing data.

Participants: Samuel Bradley (CSIRO), Aloke Phatak (CTMTDS), Gary Brown (Alcoa)
Outputs: Data Lifecycle Diagram, Data Upload and Center Collection detailed use case.

Develop a web tool to help standardise risk assessment for industry partners.

Participants: Samuel Bradley (CSIRO), John Hille (CSIRO), Joanna Sikorska (UWA), Melinda Hodkiewicz (UWA)
Outputs: DRAT Paper, Public Web Tool

Create mappings from standardised risk categories, to industry risk categories to actionable requirements and constraints. Based on industry feedback.

Participants: Samuel Bradley (CSIRO), Ryan Fraser (CSIRO), Gary Brown (Alcoa)
Outputs: Data Risk Document with mappings and actions.

Setup and develop a CKAN repository for managing and storing the data based on security and usability requirements.

Participants: Samuel Bradley (CSIRO)
Outputs: Centre Data Repository, GIT Projects (Extension enhancements & CKAN deployment)

Migrate the Prognostics Data Library to our new CKAN repository to help with development and to get some early data in prior to industry projects starting.

Participants: Samuel Bradley (CSIRO), Joanna Sikorska (UWA)
Outputs: PDL Data

FAIR

This project has laid the groundwork for receiving data from industry partners and making it as FAIR as security and confidentiality constraints will allow. This project has also detailed a data management pipeline that involves creating a broadly FAIR collection at the end of each project.

By creating a conceptual separation between *raw data from industry*, *working data* and *clean data for publishing* and applying the FAIR principles appropriately to each we are able to meet the requirements of our industry partners working with confidential data while still being able to provide FAIR data for the benefit of the research community.

See the attached FAIR assessment spreadsheet. At the end of the project we are assessing how FAIR data will be residing in the centre's internal repository and in two years time we are assessing the planned publishing of clean and de identified (where necessary) data from centre projects to a more accessible repository such as Zenodo.

Collaboration and coverage

Initially data that was previously internal to companies will be made broadly available to the CTMTDS, which is comprised of researchers from CSIRO, UWA & Curtin as well as a number of PHD Students. Members of the centre will be able to discover and collaborate on on this data.

Sustainability

The CTMTDS has funding for 5 years and all data collections and infrastructure is guaranteed for that duration. Long term sustainability of data is being solved through the additional step of publishing cleaned data to Zenodo, which is a public repository with long term sustainability as a core principle.
As CSIRO is managing the Data and the infrastructure valuable data that is unable to be published on Zendo will be transferred to the CSIRO Data Access Portal with appropriate controls in place, as is strongly encouraged by organisational policies.

Learnings

Companies are very protective of their data.

This may seem obvious in retrospect but even after agreeing to participate in the centre the companies involved didn't just throw us the keys to treasure troves of data. It has been

important to establish data management policies and infrastructure up front and to socialise these with our industry contacts.
There also seems to be a preference for a project to exist prior to related data being released (as opposed to projects being suggested based on a large pool of available data), I suspect this is to constrain the effort involved to sourcing and approving data releases as well as tying the release of data directly to a potential project outcome.

A research collaboration such as the CTMTDS functions much like an organisation.

Due to the size of the centre (50+ people) and requirements around reporting, it has faced a lot of the same needs as a medium sized business, particularly around IT. As part of the team supporting the centres technology efforts we have spend a lot of time setting up a website, an intranet, user management, authentication, servers etc.
It has been useful to be able to tie these things together with our data repository and our future plans around compute environments and code repositories but it can be quite a challenge trying to establish basic organisational infrastructure on top of solving new problems around data and development.

Research data catalogue software is still a niche.

As a software engineer I am exposed to a lot of different software, libraries, technologies, platforms etc. and each problem area has multiple well supported viable options but when it comes to managing data collections there are only a small number of choices and all of those come with pretty big caveats. Some of the main issues I found:

- None of the major cloud providers offer an out of the box solution, beyond very simple metadata.
- Open source self hosted solutions such as CKAN have a lot of moving parts and require a fair amount of configuration and some programming.
- SAAS solutions such as data.world are very few and very expensive
- Australian research bodies have fairly sparse or constrained offerings AARNET offers some basic metadata packing with cloudstor, ARDC and other help promote existing data catalogues, CSIRO has the DAP for use by CSIRO staff and collaborators.
- A good free hosted solution, Zendo, exists but does not provide robust user/team/organisation authentication and management.

Impact

What research publications or grants has/will this collection enable?

This collection once populated will be the backbone of all project and research being completed by the CTMTDS. Although most of the data will be requested and collected as a direct result of

a project, this collection will help enable the discovery of existing data to foster increased collaboration and the spawning of new projects from existing datasets.
It is hoped that the cleaned FAIR collections published at the end projects will help research on the topic of predictive maintenance beyond the life of the centre.

What new collaborations and/or communities has this project enabled?

This project has helped ensure that FAIR data principles are baked into the data collection process right from the start of the CTMTDS. This will help ensure that data is better utilised within the centre for research but will also ensure a rich source of predictive maintenance data is provided the broader research community down the track.

What new research projects or programs have been enabled?

None yet but it is expected that a FAIR data collection (enabled by the work from this project) as opposed to separately collected and stored data with enable the members of the centre to discover and integrate new data to enhance their projects or enable new ones. It also enables projects which may require data from multiple industry partners.

Will the project enable new research areas or approaches?

This is a goal of the centre and this project has contributed towards that end.

Who or what might benefit from the results of the project (industry, community, government, wider public, etc)?

The research being planned by the centre is aimed at benefiting:

- The Australian mining industry
- Other Australian Industries dependent on vehicles and machinery
- Industry workers
- The Australian Public

What is the anticipated nature of the impact, including social, economic, cultural, and environmental impacts?

- Reduced costs of Australian industry due to improved efficiency and better predictability of problems.
- Reduction in waste due to more accurate knowledge of when replacements and maintenance is necessary.
- Improved working conditions with more automation & computer assisted tasks and reduced need for fly in/fly out.
- National economic benefit due to improved industry efficiency and profitability

What is the expected extent of the impact and the time period in which it may occur?

Due to a constrained area of interest, a diverse range of projects within this area and a particular focus on providing tangible outputs that can be used by industry, it is expected that the centre will be able to create moderate to high impact within a short timeframe.
Individual project will be supported by industry meaning the researchers and engineers will have access to knowledge of industry systems and processes to go beyond pure research and develop solutions that can be integrated quickly back into industry.

Will you put in place pathways to ensure future impact?

Projects within the centre have a focus on impact and results built in. The centre provides software engineering resources and direct contact with industry to help ensure research is able to produce tangible outputs of benefit to industry.

Report prepared by: Sam Bradley CSIRO
Date: 03/10/2019